Generating Human Motion from Textual Descriptions with Discrete Representations

Year: 2023 | Citations: 491 | Authors: Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang

Abstract

In this work, we investigate a simple and must-known conditional generative framework based on Vector Quantised-Variational AutoEncoder (VQ-VAE) and Generative Pre-trained Transformer (GPT) for human motion generation from textural descriptions. We show that a simple CNN-based VQ-VAE with commonly used training recipes (EMA and Code Reset) allows us to obtain high-quality discrete representations. For GPT, we incorporate a simple corruption strategy during the training to alleviate training-testing discrepancy. Despite its simplicity, our T2M-GPT shows better performance than competitive approaches, including recent diffusion-based approaches. For example, on HumanML3D, which is currently the largest dataset, we achieve comparable performance on the consistency between text and generated motion (R-Precision), but with FID 0.116 largely outperforming MotionDiffuse of 0.630. Additionally, we conduct analyses on HumanML3D and observe that the dataset size is a limitation of our approach. Our work suggests that VQ-VAE still remains a competitive approach for human motion generation. Our implementation is available on the project page: https://mael-zys.github.io/T2M-GPT/.